GEA1000 Final
AY24/25 sem 2
github.com/mendax1234

Getting Data

1. **Census**: A method of data collection or attempts to reach **everyone** in the **population**.
2. **Bias**
   - **Selection Bias**: Due to the **researcher**'s biased selection of units.
   - **Non-response Bias**: Arises from **participants**' non-participation or non-disclosure. e.g., sending out 2000 surveys, receiving only 500 responses.
3. **Probability Sampling**
   - **Simple Random Sampling (SRS)**: Use a random generator to select each sampling unit. Any sample size n must be **equally likely to be chosen**.
     - **Advantages**: **No selection bias**. Larger sample size **reduces random errors**.
     - **Shortcoming**: Subject to **non-response** bias. Have to know the whole number of sampling units in the **sampling frame**.
   - **Systematic Sampling**:
     - **Sampling Example**:
       - The population has p sampling units in total
       - We decide our sample to have n units. We select **one unit** from every $k = \frac{p}{n}$ units;
       - From 1 to k, select a number **at random**, say r
       - With this, the sample will consist : $r, r+k, r+2k, \cdots, r+(n-1)$
     - **Advantage**: No need to know the whole sampling units in the **sampling frame**.
     - **Shortcoming**: If the **sampling frame** is **not random**, the sample **may not be representative**.
   - **Stratified Sampling**:
     - **Sampling Example**
       - The sampling frame is divided into groups called **strata**. Each **stratum** is **shares similar characteristics** but the size **may be different**.
       - SRS is then applied to each stratum to generate the whole sample.
     - **Shortcoming**: Hard to form such stratum.
   - **Cluster Sampling**:
     - **Sampling Example**
       - Divide the sampling frame into **clusters**, where **clusters** doesn't have any requirements for its inner sampling units, thus ensuring the **inner diversity**
       - Use SRS to select **a fixed number of clusters**.
       - All the sampling units from the selected clusters are then included in the overall sample.
     - **Tips**
       - Clusters can be formed by grouping students' **name**.
   - **Tips**
     - In probability sampling, every sampling unit must have a **non-zero** chance to be selected.
4. **Non-probability Sampling**
   - **Convenience Sampling**: A **researcher** chooses the sampling units **by convenience**.
   - **Volunteer Sampling**: The sampling units **volunteer** themselves into a sample, a.k.a **self-selected sampling**.
5. **Mean** $\bar{x}$: It is just **average**.
6. **Median**: Sort the data first, then find the middle value. If total number is **even**, find the **average of the middle two** values.
7. **Standard Deviation** s_x: It is computed using, $\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$. Share the **same unit** as the **numerical variable** x. In histogram, the standard deviation reverses the intuitive.
8. **Variance** = s_x^2

Operation	Mean	Median	Std Dev
Add c	$+c$	$+c$	No change
Multiply by c	$\times c$	$\times c$	$\times \lvert c \rvert$

9. **IQR**: IQR = $Q_3 - Q_1$, to find Q_3 (75% percentile) and Q_1 (25% percentile), we can use
   - **Find the median** of the total n data points.
   - **Divide the data into upper half and lower half** according to the median
     - If n is **even**, just divide normally
     - If n is **odd**, **exclude** the median from the **upper half**
   - **Find** Q_1 and Q_3
     - Q_1 is the median of the **lower half**.
     - Q_3 is the median of the **upper half**.
10. **Types of variables**
    - **Categorical Nominal Variable**: No intrinsic ordering. (e.g., "yes/no")
11. **Coefficient of variation** = $\frac{s_x}{\bar{x}}$, **no unit**
12. **Generalisability**:
    - The **sample frame** must be **greater than** the **population of interest**.
13. **Study Designs**
    - **Experimental Studies**: **Manipulate** the independent variable, e.g. **researchers** assign them **treatment** or **placebo**, to see the effect of the dependent variable.
      - **Treatment Group**: Those who receives the "treatment"
      - **Control Group**: Those who **does not receive** the "treatment", or use the **existing treatment** given that we already know the effect of **no treatment**.
      - **Placebo**: A **placebo** is a substance with **no actual effect** but is made to **look like the treatment**.
      - **Can** provide **cause-and-effect relationship** if it has features of **randomized assignment** and **blinding** (preferably double blinding).
    - **Observational Study**: **Observes** individuals and measures the variables of interest, usually **without any direct/deliberate manipulation of the variables** by the researchers.
      - We still use terms like **treatment** and **control groups**.
      - For **observational studies**, subjects assign **themselves** into either the treatment or control group.
      - **Observational studies cannot** provide cause-and-effect relationship.
      - **No selection bias** in **observational studies**.
    - **Random Assignment**: Uses chance (or probability) to **allocate objects into treatment and control groups**.
      - **Property 1**: If the number of subjects is large, by the law of probability, the subjects in the treatment and control groups will tend to be **similar in all aspects** (like **have similar characteristics**).
      - **Property 2**: When performing random assignment, the size of **treatment group** and **control group** does not need to be **the same**.
    - **Blinding**
      - **Double-blinding** doesn't ensure the **generalisability**.

Categorical Data Analysis

1. **Rate**: It is calculated as the ratio of the number of observations in a **given category** (a.k.a **target**) to the **total** number of observations (a.k.a **population**).
   - **Tips**: Regarding rate problem, always find your **target** and **population**. Then rate $=$ target $\div$ population
2. **2x2 contingency table**: **Dependent Variables** at **columns**, **Independent variables** at **rows**. For example, **treatment** is independent variable, **outcome** is the dependent variable.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

3. **Marginal rate**: rate(A), **A** is the **target**, the **population** is the total by default.
4. **Conditional rate**: rate($A \mid B$), our **target** will be the number of A under the condition B, our **population** will be the total number which satisfies the condition B.
5. **Joint Rate**: rate($A \cap B$) our **target** will be the **intersection/and**, the **population** will be the whole population by default.
6. **Association**: Describes a relationship between two **categorical variables**. Association $\neq$ causation.
   - **Positive association**: rate($A \mid B$) $>$ rate($A \mid NB$). Meaning: the presence of A **when B is present** is **stronger** compared to when B is absent.
   - **Negative association**: rate($A \mid B$) $<$ rate($A \mid NB$). Meaning: the presence of A **when B is present** is **weaker** compared to when B is absent.

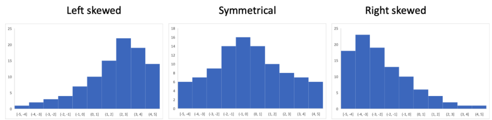
Establishing association	
Positive association between A and B: (any of the following)	Negative association between A and B: (any of the following)
rate($A \mid B$) > rate($A \mid NB$)	rate($A \mid B$) < rate($A \mid NB$)
rate($B \mid A$) > rate($B \mid NA$)	rate($B \mid A$) < rate($B \mid NA$)
rate($NA \mid NB$) > rate($NA \mid B$)	rate($NA \mid NB$) < rate($NA \mid B$)
rate($NB \mid NA$) > rate($NB \mid A$)	rate($NB \mid NA$) < rate($NB \mid A$)

   - **Tips**: Always find what **A** is and what **B** is.
7. **Two rules on rates**
   - **Symmetric Rule**
     - **Part 1**: rate($A \mid B$) $>$ rate($A \mid NB$) $\Leftrightarrow$ rate($B \mid A$) $>$ rate($B \mid NA$)
     - **Part 2**: rate($A \mid B$) $<$ rate($A \mid NB$) $\Leftrightarrow$ rate($B \mid A$) $<$ rate($B \mid NA$)
     - **Part 3**: rate($A \mid B$) $=$ rate($A \mid NB$) $\Leftrightarrow$ rate($B \mid A$) $=$ rate($B \mid NA$)
   - **Basic Rule on Rates**: The overall rate(A) will always lie between rate($A \mid B$) and rate($A \mid NB$)
     - **Consequence 1**: The closer rate(B) is to 100%, the closer rate(A) is to rate($A \mid B$)
     - **Consequence 2**: If rate(B) $= 50\%$, then rate(A) $= \frac{1}{2}$[rate($A \mid B$) $+$ rate($A \mid NB$)]
     - **Consequence 3**: If rate($A \mid B$) $=$ rate($A \mid NB$), then rate(A) $=$ rate($A \mid B$) $=$ rate($A \mid NB$)
     - **Tips**: The bounds for the interval will change to whatever is smaller and bigger.
8. **Simpson's Paradox**: If we divide the whole population into several subgroups, the trend that appears in **more than half of the subgroups** of data may **disappears or reverses** when the subgroups are combined, this is called **Simpson's Paradox**.
   - The appearance of **simpson's paradox** implies the variable we use to slice is a **confounder**.
9. **Confounder**: A third variable that is **associated** with **both the independent and dependent variables**.
   - The appearance of a **confounder** does not necessarily imply a **simpson's paradox**.
   - Remove **one of the associations** is enough to remove the confounding variable.
10. **Tips**
    - When building examples for questions regarding **median or mean**, be very careful! Use exhaustive thinking!
    - rate(A | B) is **not equal to** rate(B | A)!
    - See words like **can**, build an example to prove!

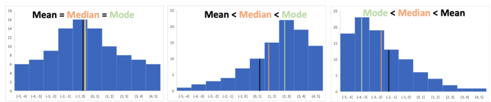
Dealing with Numerical Data

Univarite EDA

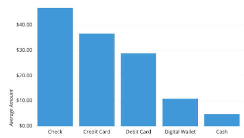
1. **Skewness**: Used only in **unimodal distribution**



And its central tendency is as follows



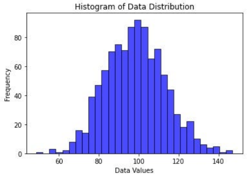
2. **Range** = the largest data point - the smallest data point
3. **Outlier**: If the value is either **greater than** $Q_3 + 1.5 \times$ IQR or **less than** $Q_1 - 1.5 \times$ IQR (Equal is **not include**! a.k.a, the range is **open**, not **bounded**)
4. **Bar Chart**



   - Used to display data for **categorical** variables (**not numerical variables**)
   - Typically have gaps between each bar

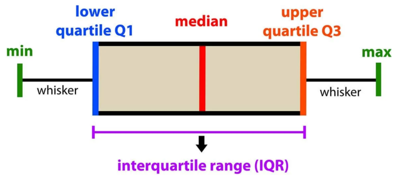
- The order of bars can be **rearranged freely**.

5. **Histogram**:



- The width of each rectangle is called **bin width**.
- The number of **data points** we have in a **data set** is better shown in a **histogram** than in a **boxplot**.

6. **Boxplot**:

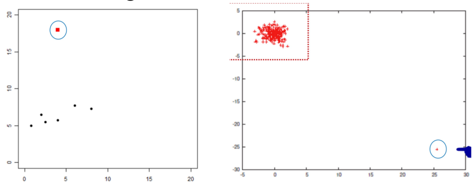


- Sometimes the $\times$ near the **median** line is used to indicate the **mean**.
- **Outliers** are shown as **dots**, so the number of dots indicates how many outliers there are. Thus, **boxplots** are better at identifying outliers than **histograms**.
- **Identical boxplots do not imply** the same number of data points or the same standard deviation.

7. **Tips**
   - When constructing the counterexample, the value in a data set can be **negative** unless restrictions are specified elsewhere.
   - In **scatter plot**, an **outlier** is a data point that deviates significantly from the **overall pattern** (the cluster of points) or **trend of the other points** (the regression line of points)

Bivariate EDA

1. **Direction**: describes the **relationship** between two variables. Can be **positive, negative or neither**.
2. **Form**: Describe the overall **shape** of a scatter plot. It is classified into **linear** and **non-linear** (which may include quadratic or exponential patterns)
3. **Correlation Coefficient**: A measure of **linear** association between two numerical variables. Always between -1 and 1. The **sign** of r reveals the direction, while the **magnitude** (how close r is to 1 or -1) indicates the **strength** of the association.
   - **Three properties related to** r: r is **not** affected by 1)**interchanging** the x and y variables, 2)**adding** a number to **all** values of a variable and 3)**multiplying** a **positive** number to **all** values of a variable.
   - **How removing outliers will affect** r

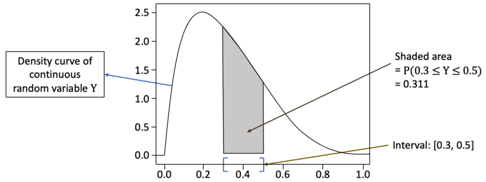


   - In the Left figure, removing outlier will **increase** the **strength**
   - In the right figure, removing outlier will **decrease** the **strength**. e.g., r decreases from 0.75 to 0.01.
   - Removing the outlier may also **not change** the r.
   - r has **sign**, thus **not having the same** change as **strength**, be careful!
4. **Ecological Fallacy**: Use **aggregate level** correlation to conclude **individual level correlation**.
5. **Atomistic Fallacy**: The reverse of Ecological Fallacy.
6. **Linear Regression**:
   - The linear regression line **always** pass through the average point $(\bar{x}, \bar{y})$.
   - Make prediction only **within the range of independent variable**.
   - Removing the outlier may **increase, decrease or not change** the **slope** of the linear regression line.
7. **Tips**
   - **Association** is **not causation**.
   - The correlation coefficient r does not tell anything about **non-linear** relationship. While r for a non-linear relationship can be small, its relationship may be actually **strong**.
   - Correlation coefficient r has the **same sign** with the **slope of the linear regression line**.

Statistical Inference

Probability

1. **Basic Terms**:
   - **Sample space**: The collection of **all possible outcomes** of a probability experiment. e.g. [HH, TT, HT, TH]
   - **Event**: A **sub-collection** of the sample space is called an **event**. (Think it as **subset**)
   - **Outcome**: It is exactly the event of **one element** in the sample space.
2. **Conditional Probability**: $P(E \mid F) = \frac{P(E \cap F)}{P(F)}$, which means the probability of E to happen **given that** F happens.
3. **Prosecutor's fallacy**: The mistake of confusing $P(A \mid B)$ as $P(B \mid A)$
4. **Independent Event**: We say two events A and B are **independent** if and only if $P(A \cap B) = P(A) \times P(B)$
5. **Mutually Exclusive Event**: Two events **cannot** happen together, which means $P(E \cap F) = 0$. E and F are mutually exclusive if and only if $P(E \cup F) = P(E) + P(F)$
6. **Conditional Independency**: We say that two events A and B are **conditionally independent** given an event C if $P(A \cap B \mid C) = P(A \mid C) \times P(B \mid C)$
7. **Law of total probability**: If E, F, G are events from the same sample space S such that 1) E and F are mutually exclusive, and 2)$E \cup F = S$, then $P(G) = P(G \cap E) + P(G \cap F) = P(G \mid E) \times P(E) + P(G \mid F) \times P(F)$
8. **Conjunction Fallacy**: You think that the probability of two things happening together is **greater** than one thing happens. But **it is not true**!
9. **Base rate fallacy**: The mistake that only **sensitivity and specificity** are given, but **base rate** is ignored.
   - **Sensitivity**: This is same as **true positive rate**. e.g. $P(\text{Test positive} \mid \text{Individual is infected})$
   - **Specificity**: This is same as **true negative rate**. e.g. $P(\text{Test negative} \mid \text{Individual is not infected})$
   - **Base rate**: e.g. The infection rate $P(\text{Individual is infected})$
   - Regarding this kind of question, always **build a 2x2 contingency table**. Always start from the **conditional base rate** or the **base rate**. Then use **sensitivity** and **specificity**.
10. **Random Variable**



x-axis is the possible value for the random variable Y, y-axis is the **probability density**. This graphs means $P(0.3 \le Y \le 0.5) = 0.311$

Confidence Interval

1. **For population proportion**
   - **Formula**: $p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$, where p^* = sample proportion, z^* = "z-value" from standard normal distribution, n = sample size
   - z^* **value**: 90% CI, $z^* = 1.645$, 95%CI, $z^* = 1.96$. More confident, z^* bigger.
2. **For population mean**
   - **Formula**: $\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$, where $\bar{x}$ = sample mean, t^* = "t-value" from t-distribution, s = sample standard deviation, n = sample size
   - t^* **value**: More confident, t^* bigger.
3. **Margin of error**: The term after $\pm$ in the formula.
4. **Interpretation**: If **many simple random samples** of the same size are taken, and a **confidence level** is constructed for each of them, then about **95% of the CI** constructed would contain the **population parameter**. But every CI will contain its corresponding **sample population parameter**.
5. **Properties of CI**
   - The **larger the sample size** n, the **smaller the random error** (a.k.a margin error).
   - The **higher the confidence level** at which the CI is constructed (a.k.a the larger the z^* or t^*), **the wider the CI**.
6. **Tips**
   - Given a CI with the **same sample**, always calculate the **sample population parameter** and the **margin of error**.
   - CI is a way to **quantify** the random of error.
   - CI constructed from **sample** is **likely** to include the population parameter. But if CI is constructed from **population**, it confirms to contain the **population parameter**.

Hypothesis Testing

1. **Definition**: A **hypothesis test** is a statistical inference method used to decide if the data from a random or even more extreme is **sufficient** to support a particular hypothesis about a population.
2. **Two cases**:
   - a population parameter is x (denoted as Case 1)
   - in the population, 2 categorical variables A and B are associated with each other. (denoted as Case 2)
3. **Null hypothesis**
   - In case 1, it says population parameter p equals to a specific value.
   - In case 2, it means there is **no association** (Independence) between the two categorical variables.
4. **Alternative hypothesis**
   - In case 1, it says population parameter $p \neq$ a specific value.
   - In case 2, it means there is **an association** (Dependence) between the two categorical variables.
5. **Five steps of hypothesis testing**
   - Identify the question and state **the null hypothesis** and **alternative *hypothesis***.
   - Set the **significance level** of our test. It is often set at 5%, can be 1% and 10% also.
   - Using our sample, we find the relevant sample statistic. This means calculating the population parameter we want but using the **sample data**.
   - With the sample statistic and the hypothesis, we can calculate the p-value.
   - Make a conclusion of the hypothesis test.
     - If the p-value is $\le$ the significance level (e.g., 0.05), you **reject the null hypothesis** and say there is evidence for the alternative hypothesis.
     - Otherwise, you **fail to reject the null hypothesis**, we don't have enough evidence to support the alternative. (You don't "accept" the null.)
6. **The meaning of p-value**: The probability of obtaining a result **as extreme** or **more extreme** than our observation (the value calculated using the sample data) **in the direction of alternative hypothesis** (This means following the direction!), assuming **the null hypothesis is true**.
7. **Tips**
   - For example, if your suspect is biased **against** heads, your **alternative hypothesis** should be $P(\text{Heads}) < 0.5$
   - **Steps to find other observations need to be considered when calculating p-value**
     - Know the direction of your **alternative hypothesis**
     - We need the observations **starting from the initial observation** and **follow the direction** in the above step. Just follow.
     - **Decimal points**: the digits after the **.** e.g. 2 decimal points, 3.14
     - **Significant points**: the number of digits starting from first non-zero. e.g., 2 significant points, 0.0045